The peakpicking algorithm is coming along well and shows a lot of promise for being able to automatically identify peaks of interest and significantly improve our ability to confidently isolate important differences in any set of mass-spec files. Quick demo of progress and thoughts below.

Motivation: Identifying real peaks amidst the noisy mass spectra is a huge bottleneck in our pipeline, often requiring manual curation and false positive removal. But it doesn’t have to be! By introducing a metric of peak *quality* alongside each peak, we can focus our energy on the important peaks that are more likely to correspond to real compounds OR set a lower threshold for peak quality and pass those straight to the rest of the pipeline (QC/BMIS, hypothesis testing, figure creation). However, no other peakpicking algorithms have this kind of confidence assessment built in, meaning that I’m writing my own. This also gives me much finer control over the specific parameters that we find important.

Currently programmed and functioning well: EIC construction, peakpicking, and isotope identification

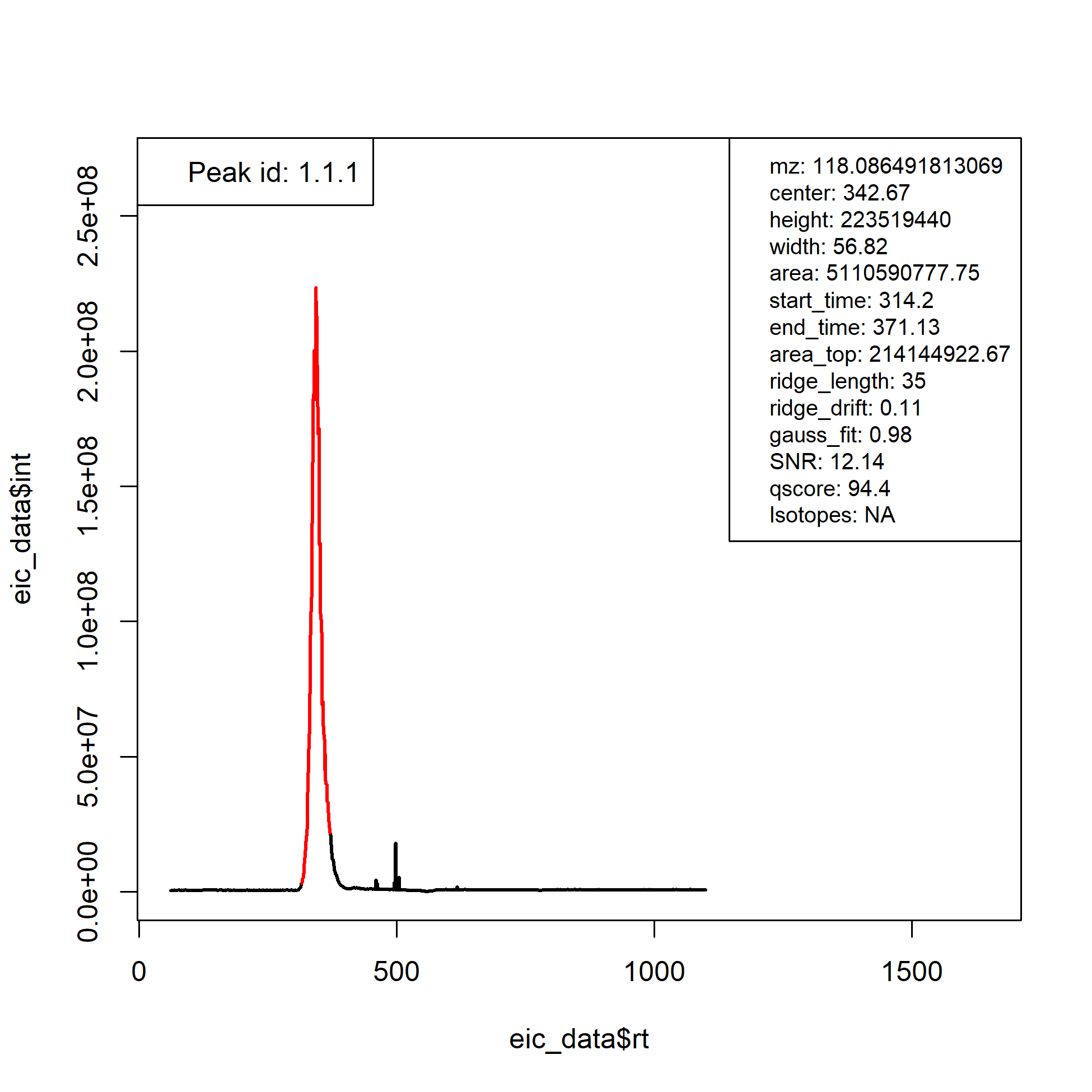
Next step: multi-file handling (in parallel?) and retention time correction

Next next step: calibrating peak quality metrics against real-world data and the human eye

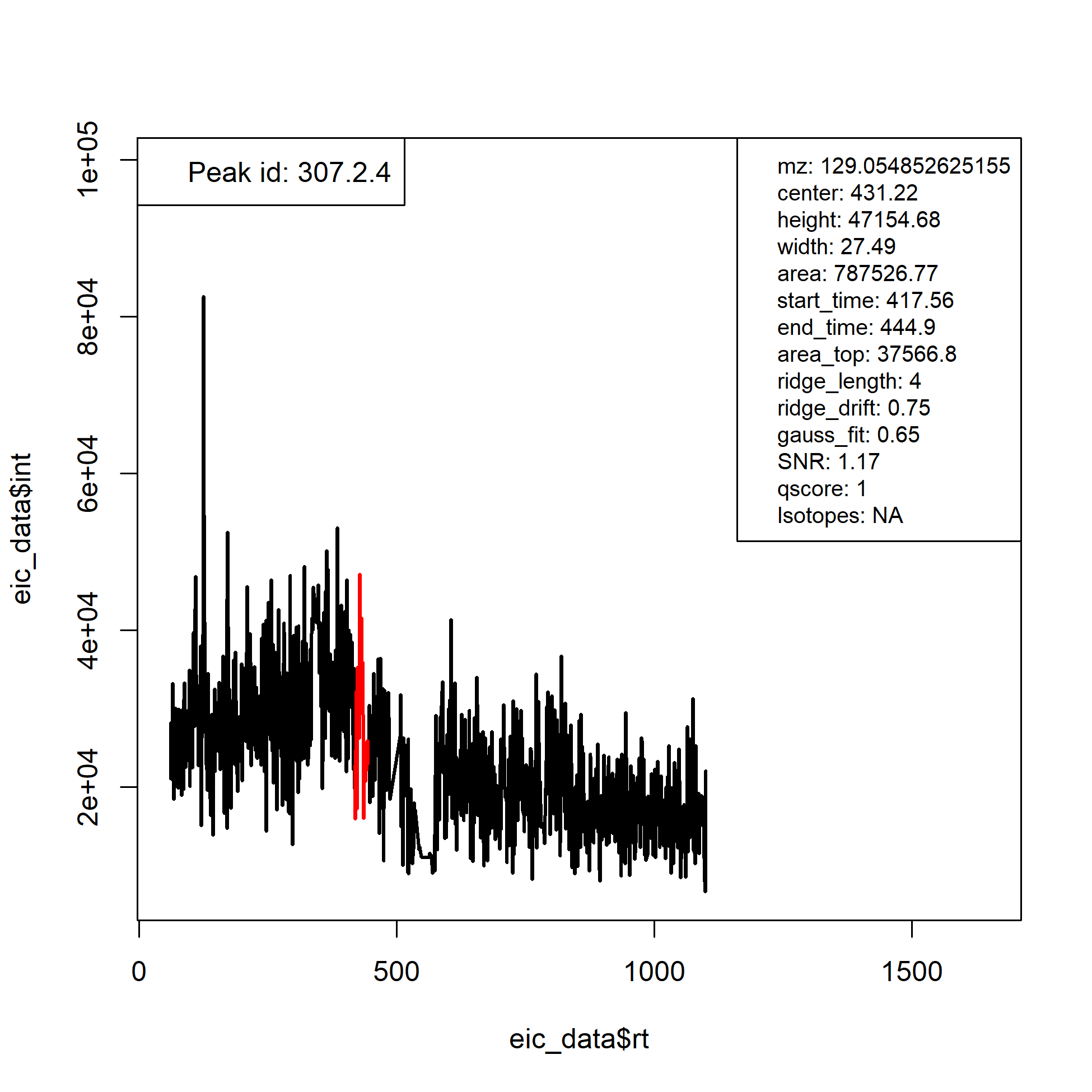
Peak picking quality!

Good peaks will have a high signal to noise (SNR), Gaussian fit close to 1 (gauss\_fit), and are generally tall (peak\_height). Combining these metrics (and others? see below) allows us to come up with an overall quality metric (qscore). However, this is merely a relative quality score until we’re able to actually able to compare a bunch of peaks as identified by human eye vs the computer – then we’ll be able to say things like “90% of the time, a human would identify this as a real peak” or “1% of the time, a human would identify this as a peak” by creating a predictive logistic model for each of the peak metrics.

## Good peak (with high qscore metric = 94.4):



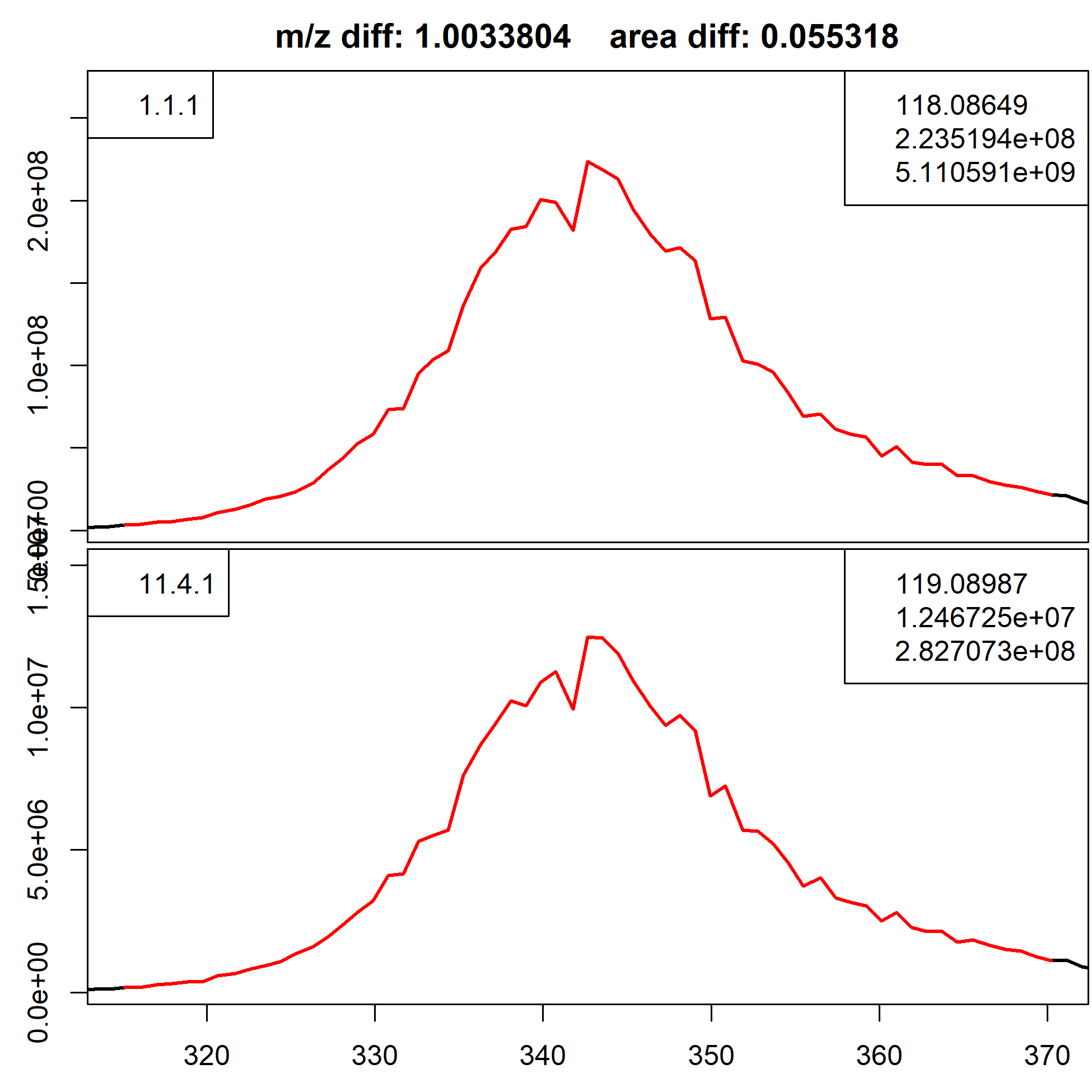
## Bad peak (with low qscore metric = ~1)



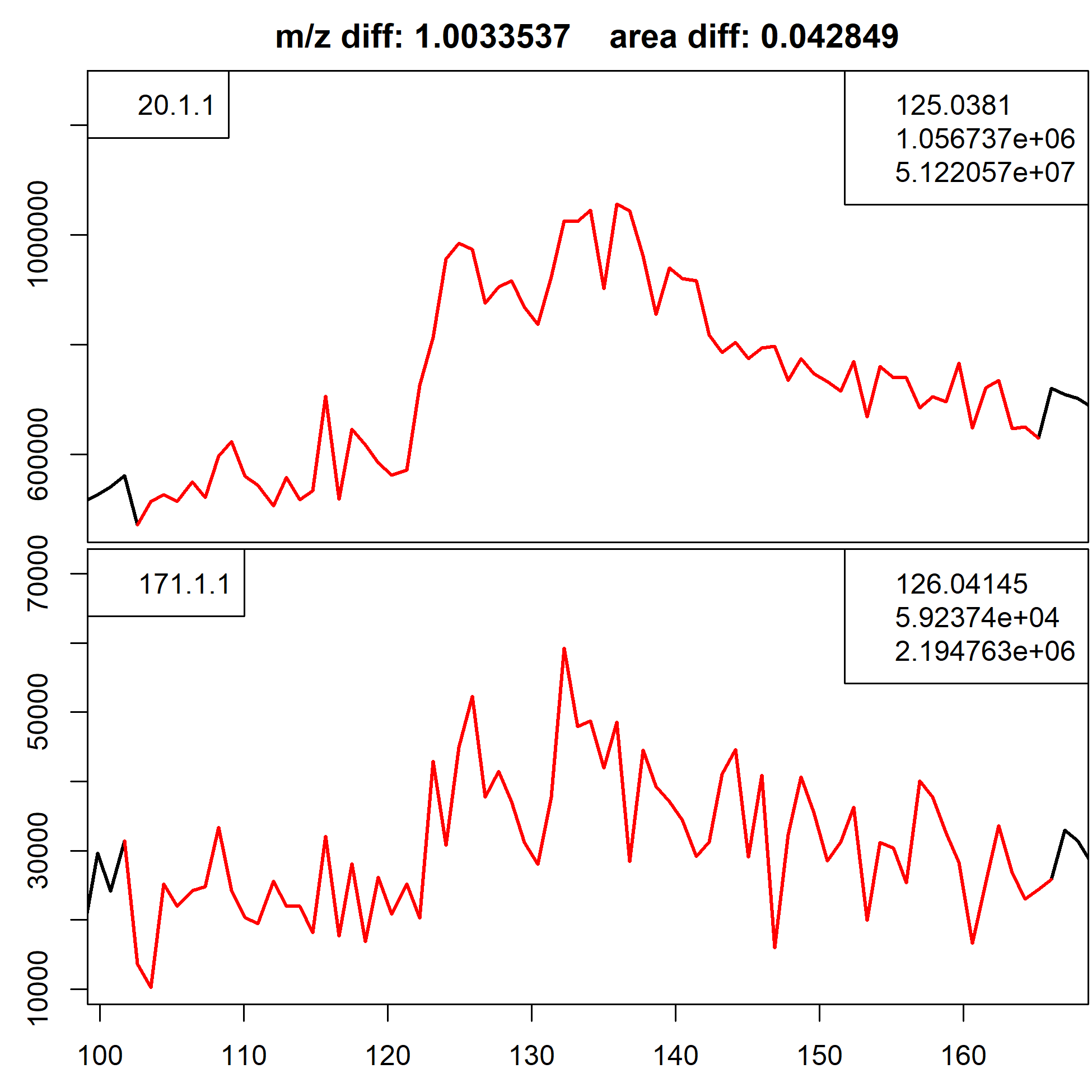
Isotope picking quality!

Another good clue that a peak is truly a peak is whether or not it has isotopes associated with it. So we search for isotopes within the instrumental error of the C13 value for each peak, and can again assess the likelihood that a peak is truly an isotope by 1) the m/z match (close to 1.003355, for C13-C12), 2) how close the peak centers are to each other, and how well the peak shapes correlate to each other. For example:

## Good isotope match!

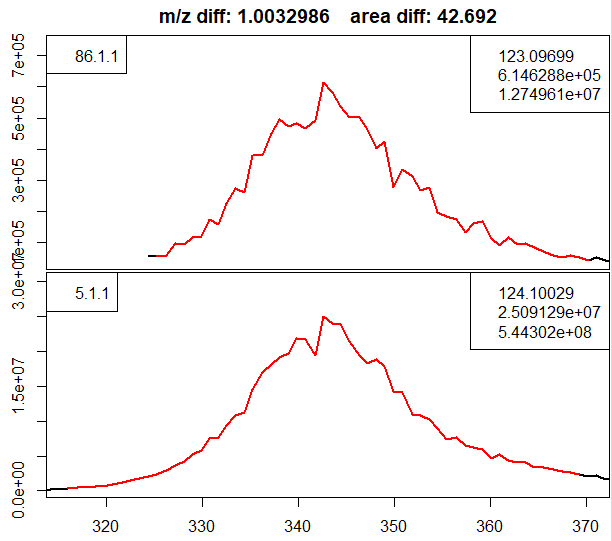


## Worse isotope match (not *terrible* and it’s probably still an isotope, but not great either)



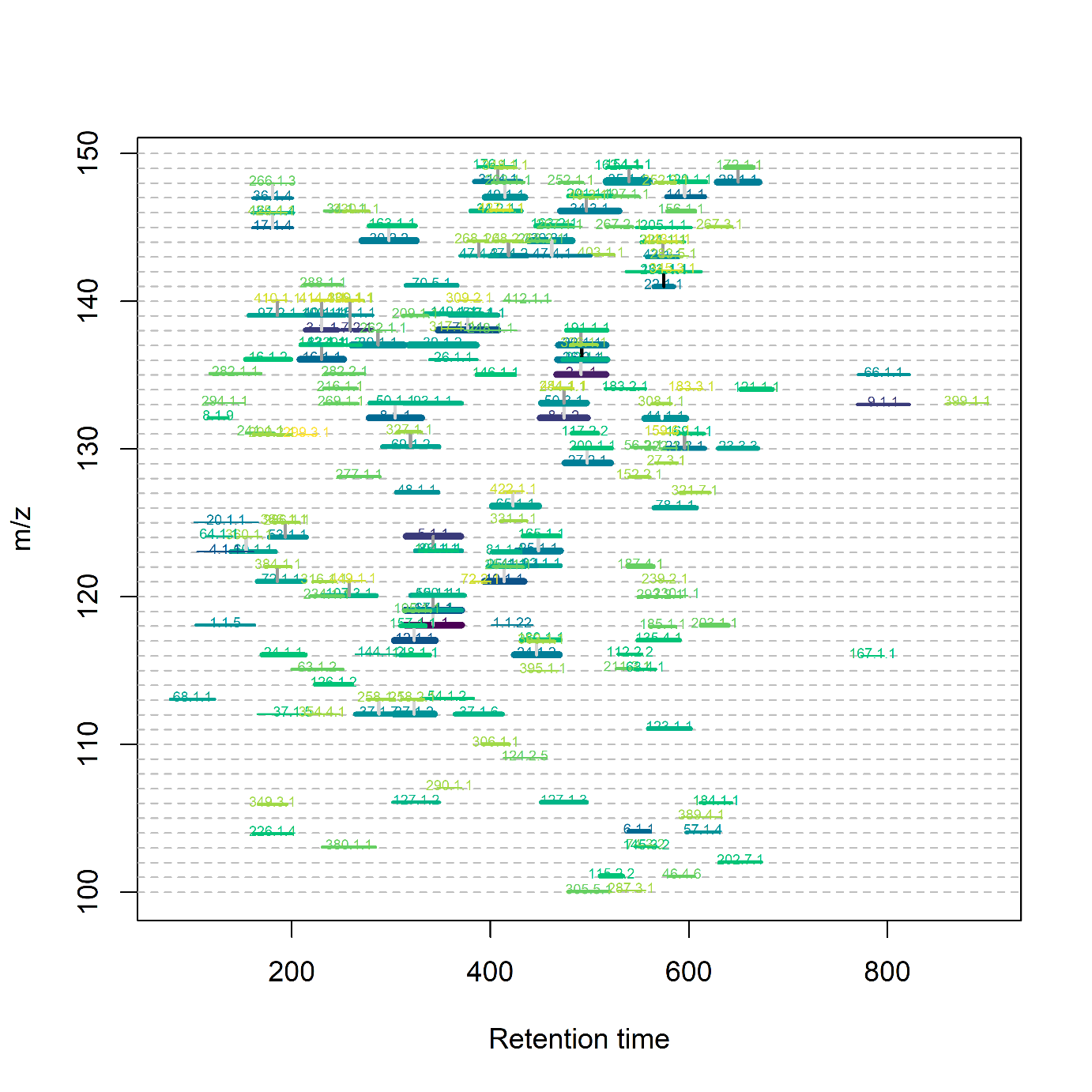
Interestingly, the difference in areas between these peaks shouldn’t be 0.011 – the abundance of C13 relative to C12. Rather, because each compound may contain multiple carbons, it’s actually the binomial probability of each carbon in the molecule being C13. This actually tells us how many carbons are in the molecule, which enormously constrains the molecular formula and structural possibilities when we’re doing a database lookup or in-silico simulation to identify it. For betaine in the “Good isotope” above, the binomial formula would predict a difference in areas of 0.055318 is pretty close to the predicted binomial distribution for 5 carbons (0.05262). And, sure enough, betaine has 5 carbons. Credit to Katherine for figuring that one out!

So for each peak we find, we can do a quick search for isotopes for each one and the probability of those actually being isotopes. This can then feed back into our initial peak confidence estimate and raise or lower the likelihood based on whether or not it’s actually a peak. This was accidentally validated by me finding one of our C13 standards in the sample – the peak area difference estimate is wildly different, but that’s because we’ve added the C13 itself. Credit to Angie for figuring that one out!



Internal Standard HILIC Betaine

In the above feature for our internal standard Betaine, tagged with 5 C13 carbons and one N15 carbon, we get a relative peak area of 42. Useful for identifying isotopes!

So the final output of the algorithm is currently visualized below (also attached to the email because quality is low here)

Good peaks are darker in color, where low-quality ones are yellowish. The length of the line is the retention time, and the thickness is peak area. Grey lines connecting two peaks indicate an isotopic connection, with the darkness of the connection based on the likelihood of it actually being an isotope as outlined above.

What’s great about the algorithm when it’s finished is that it’s fully general – I can run HILIC or lipid, positive or negative, with minimal changes to the settings. I expect to build this into the getMetlin and Rdisop packages to continue to improve our peak confidence and hopefully find some new important things!